# The Stata Journal

The *Stata Journal* publishes reviewed papers together with shorter notes or comments, regular columns, book reviews, and other material of interest to Stata users. Examples of the types of papers include 1) expository papers that link the use of Stata commands or programs to associated principles, such as those that will serve as tutorials for users first encountering a new field of statistics or a major new technique; 2) papers that go "beyond the Stata manual" in explaining key features or uses of Stata that are of interest to intermediate or advanced users of Stata; 3) papers that discuss new commands or Stata programs of interest either to a wide spectrum of users (e.g., in data management or graphics) or to some large segment of Stata users (e.g., in survey statistics, survival analysis, panel analysis, or limited dependent variable modeling); 4) papers analyzing the statistical properties of new or existing estimators and tests in Stata; 5) papers that could be of interest or usefulness to researchers, especially in fields that are of practical importance but are not often included in texts or other journals, such as the use of Stata in managing datasets, especially large datasets, with advice from hard-won experience; and 6) papers of interest to those who teach, including Stata with topics such as extended examples of techniques and interpretation of results, simulations of statistical concepts, and overviews of subject areas.

The *Stata Journal* is indexed and abstracted by *CompuMath Citation Index*, *Current Contents/Social and Behavioral Sciences*, *RePEc: Research Papers in Economics*, *Science Citation Index Expanded* (also known as *SciSearch*, *Scopus*, and *Social Sciences Citation Index*.

For more information on the *Stata Journal*, including information for authors, see the webpage

http://www.stata-journal.com

**Subscriptions** are available from StataCorp, 4905 Lakeway Drive, College Station, Texas 77845, telephone 979-696-4600 or 800-STATA-PC, fax 979-696-4601, or online at

http://www.stata.com/bookstore/sj.html

**Subscription rates** listed below include both a printed and an electronic copy unless otherwise mentioned.

| U.S. and Canada | | Elsewhere | |
|---|---|---|---|
| **Printed & electronic** | | **Printed & electronic** | |
| 1-year subscription | $ 98 | 1-year subscription | $138 |
| 2-year subscription | $165 | 2-year subscription | $245 |
| 3-year subscription | $225 | 3-year subscription | $345 |
| 1-year student subscription | $ 75 | 1-year student subscription | $ 99 |
| 1-year university library subscription | $125 | 1-year university library subscription | $165 |
| 2-year university library subscription | $215 | 2-year university library subscription | $295 |
| 3-year university library subscription | $315 | 3-year university library subscription | $435 |
| 1-year institutional subscription | $245 | 1-year institutional subscription | $285 |
| 2-year institutional subscription | $445 | 2-year institutional subscription | $525 |
| 3-year institutional subscription | $645 | 3-year institutional subscription | $765 |
| **Electronic only** | | **Electronic only** | |
| 1-year subscription | $ 75 | 1-year subscription | $ 75 |
| 2-year subscription | $125 | 2-year subscription | $125 |
| 3-year subscription | $165 | 3-year subscription | $165 |
| 1-year student subscription | $ 45 | 1-year student subscription | $ 45 |

Back issues of the *Stata Journal* may be ordered online at

http://www.stata.com/bookstore/sjj.html

Individual articles three or more years old may be accessed online without charge. More recent articles may be ordered online.

http://www.stata-journal.com/archives.html

The *Stata Journal* is published quarterly by the Stata Press, College Station, Texas, USA.

Address changes should be sent to the *Stata Journal*, StataCorp, 4905 Lakeway Drive, College Station, TX 77845, USA, or emailed to sj@stata.com.

Users of any of the software, ideas, data, or other materials published in the *Stata Journal* or the supporting files understand that such use is made without warranty of any kind, by either the *Stata Journal*, the author, or StataCorp. In particular, there is no warranty of fitness of purpose or merchantability, nor for special, incidental, or consequential damages such as loss of profits. The purpose of the *Stata Journal* is to promote free communication among Stata users.

The *Stata Journal* (ISSN 1536-867X) is a publication of Stata Press. Stata, stata, Stata Press, Mata, mata, and NetCourse are registered trademarks of StataCorp LP.

The Stata Journal (2013)
**13**, Number 2, pp. 398–400

# Speaking Stata: Creating and varying box plots: Correction

Nicholas J. Cox
Department of Geography
Durham University
Durham, UK
n.j.cox@durham.ac.uk

A previous article (Cox 2009) discussed the creation of box plots from first principles, particularly when a box plot is desired that `graph box` or `graph hbox` cannot provide.

This update reports and corrects an error in my code given in that article. The problems are centered on page 484. The question is how to calculate the positions of the ends of the so-called whiskers.

To make this more concrete, the article's example starts with

```
. sysuse lifeexp
. egen upq = pctile(lexp), by(region) p(75)
. egen loq = pctile(lexp), by(region) p(25)
. generate iqr = upq - loq
```

and that holds good.

Given interquartile range (IQR), the position of the end of the upper whisker is that of the largest value not greater than the upper quartile + 1.5 IQR. Similarly, the position of the end of the lower whisker is that of the smallest value not less than the lower quartile − 1.5 IQR.

The problem lines are on page 484:

```
. egen upper = max(min(lexp, upq + 1.5 * iqr)), by(region)
. egen lower = min(max(lexp, loq - 1.5 * iqr)), by(region)
```

This code works correctly if there are no values beyond where the whiskers should end. Otherwise, it yields upper quartile + 1.5 IQR as the position of the upper whisker, but this position will be correct only if there are values equal to that. Commonly, that position will be too high. A similar problem applies to the lower whisker, which commonly will be too low.

More careful code might be

```
. egen upper2 = max(lexp / (lexp < upq + 1.5 * iqr)), by(region)
. egen lower2 = min(lexp / (lexp > loq - 1.5 * iqr)), by(region)
```

That division `/` may look odd if you have not seen it before in similar examples. But it is very like a common kind of conditional notation often seen, max(*argument* | *condition*)

or

min(*argument* | *condition*)

where we seek the maximum or minimum of some argument, restricting attention to cases in which a specified condition is satisfied, or true.

The connection is given in this way. Divide an argument by a logical expression that evaluates to 1 when the expression is true and 0 otherwise. The result is the argument remains unchanged on division by 1 but evaluates as missing on division by 0. In any context where Stata ignores missings, that is what is wanted. True cases are included in the computation, and false cases are excluded.

This "divide by zero" trick appears not to be widely known. There was some publicity within a later article (Cox 2011).

Turning back to the box plots, we will see what the difference is in our example.

```
.  tabdisp region, c(upper upper2 lower lower2)

─────────────────────────────────────────────────────────────
       Region │      upper      upper2       lower      lower2
──────────────┼──────────────────────────────────────────────
Eur & C.Asia  │         79          79          65          65
         N.A. │         79          79        58.5          64
         S.A. │         75          75          63          67
─────────────────────────────────────────────────────────────
```

Here `upper2` and `lower2` are from the more careful code just given, and `upper` and `lower` are from the code in the 2009 column. The results can be the same but need not be.

Checking Stata's own box plot

```
. graph box lexp, over(region) yli(75 79 64 65 67)
```

shows consistency with the corrected code.

Thanks to Sheena G. Sullivan, UCLA, who identified the problem on Statalist (http://www.stata.com/statalist/archive/2013-03/msg00906.html).

## 1 References

Cox, N. J. 2009. Speaking Stata: Creating and varying box plots. *Stata Journal* 9: 478–496.

———. 2011. Speaking Stata: Compared with ... *Stata Journal* 11: 305–314.

### About the author

Nicholas Cox is a statistically minded geographer at Durham University. He contributes talks, postings, FAQs, and programs to the Stata user community. He has also coauthored 15 commands in official Stata. He wrote several inserts in the *Stata Technical Bulletin* and is an editor of the *Stata Journal*.